A Direct Linguistic Induction Method For Systems

Juan Moreno García
E. U. de Ingeniería Técnica Industrial
Universidad de Castilla-La Mancha
Avda Carlos III, s/n, Toledo
e-mail: jmgarcia@iti-to.uclm.es

Luis Jiménez Linares
Escuela Superior de Informática
Universidad de Castilla-La Mancha
Ronda de Calatrava, 5, Ciudad Real
e-mail: ljimenez@inf-cr.uclm.es

Abstract

The aim of this paper is presented a method for obtaining a linguistic model that reflects the behavior of a combined well-known data. The method is based on the technique of successive division of the input space (as CART[1] and ID3[4]). The obtained rules have linguistic variables as antecedent and consequent. So, incorporating the concept of linguistic intervals (as disjunctions of linguistic labels) the methods ID3 and CART are generalized for working directly with linguistic variables defined a priori.

Keywords: Linguistic Induction and Model.

1 Introduction

A linguistic induction method was develops for getting the linguistic model, this method needs a set of examples E and a set of sort linguistic labels [6] for each one of input variables and the output variable. The set E is formed by elements with the structure $e_i=(x_1^i, ..., x_m^i, y^i, t^i)$, where x_j^i is a real number defined in the domain of input variable X_j, y^i is a real number defined in the domain of output variable Y, and t is the time of the example. For every one of the input variables X_j a set of linguistic labels SA_j is defined. The set SA_j is represented as $SA_j=\{SA_j^1, SA_j^2, ..., SA_j^{i_j}\}$, where in SA_j^i i is the position of the linguistic label and j is the number of the input linguistic variable, and i_j is the number of linguistic labels in SA_j. A set of linguistic labels SC is defined for the output variable Y. The set SC has the structure: $SC=\{SC^1, SC^2, ..., SC^{i_y}\}$, where i is the position of the linguistic label in SC^i, and i_y is the number of linguistic labels in SC. The antecedent of every rule is a conjunction of disjunctions of linguistic labels, and the consequent is a linguistic label of the set SC.

2 Induction Algorithm

The scheme of the induction algorithm is:

M_{act}={IF (x_1 is SA_1^i OR... OR x_1 is $SA_1^{i_1}$) AND ..
AND (x_m is SA_m^1 OR ... OR x_m is $SA_m^{i_m}$) THEN c is SC^{i_y} }
WHILE (SomeRuleCanBeDivided) DO
R=GetRule(M_{act}); //Get a rule from M_{act} that can be divided
SPAR=Split(R); //SPAR is a Set of Pair of Antecedents
PR=SelectPairOfAntecedent(SPAR); //Select a pair
Simplify(PR); //The antecedents of PR are simplified
NR=CreateRules(PR); // Obtain the New Rules from PR
M_{act}= M_{act}-R+NR_1+NR_2;//Update M_{act} with the rules of NR
END_WHILE

In the algorithm, M_{act} is a linguistic model, R is the selected rule from M_{act}, SPAR is a set that has pairs of set of antecedent of rules (pair of set of linguistic intervals) as elements, PR is a pair chosen from SPAR, and NR is a pair of rules obtained from PR. Initially, the algorithm creates the first model M_{act} with a single rule: its antecedent has all possible linguistic labels for all input variables, and the linguistic label of the consequent is calculated (section 2.4). Next, a loop is repeated while some rule can be divided, the first sentence selects a rule R of M_{act} that can be divided (section 2.1). Then the antecedent of R is divided in pairs of set of linguistic intervals, which are introduced in the set SPAR (section 2.1). After the best pair in SPAR is selected using an evaluation function (section 2.4), the selected pair is introduced in PR. The antecedents of PR can be simplified (section 2.3), and two rules are made using simplified PR, for this, the linguistic label of consequent of the new rules is calculating. Finally, the original rule is removed from M_{act}, and the rules of NR are introduced in M_{act}. This process continues until that any rule can't be divided. An important feature of the algorithm is that during all process a set of examples E_R is associated for every antecedent of a rule R. This set is a sub-set of the set

of examples E, and is compounded by the examples of E that verify the antecedent of the rule.

2.1 Selecting a rule and Splitting its antecedent

A rule R must be selected for splitting its antecedent. R must verify a condition for could be selected: for some of the feasible pair of antecedents of rules R1 and R2 obtained when splitting the rule R, the set of examples E_{R1} and E_{R2} of the antecedent of R1 and R2 respectively must have one or plus elements, this condition will be represented formally soon.

For getting SPAR, for each disjunction of the antecedent of the linguistic rule R is made a combination of the successive linguistic labels named linguistic interval. For example, lets $(SA_j^1 \vee SA_j^2 \vee SA_j^3 \vee SA_j^4)$ the disjunction j of the antecedent of the rule R, the figure 1 represents the linguistic intervals obtained.

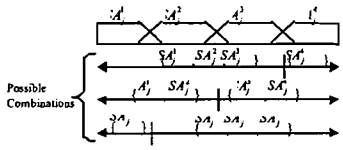

Figure 1. Combinations of a linguistic interval.

The linguistic labels of the disjunction (linguistic interval too) are grouped in the possible pairs of linguistic interval. If a linguistic interval has n linguistic labels, there are $n-1$ possible pairs. Each pair of linguistic intervals is used for making the antecedents of a pair of rules. For example, lets us suppose that R is the selected rule for dividing, and has the form:

IF (x_1 is SA_1^2 OR x_1 is SA_1^3) AND (x_2 is SA_2^1 OR x_2 is SA_2^2 OR x_2 is SA_2^3 OR x_2 is SA_2^4) AND (x_3 is SA_3^4 OR x_3 is SA_3^5) THEN y is SC^3

Using the method show in the figure 1 the possible linguistic intervals of the disjunction 2 of the R are:

Pair 1: $\{SA_2^1\}, \{SA_2^2, SA_2^3, SA_2^4\}$
Pair 2: $\{SA_2^1, SA_2^2\}, \{SA_2^3, SA_2^4\}$
Pair 3: $\{SA_2^1, SA_2^2, SA_2^3\}, \{SA_2^4\}$

This process is realized with all disjunctions of the antecedent of R. So, if the rules have m disjunctions in the antecedent, the set SPAR has $\sum_{j=1}^{m}(t_j - 1)$ pair of linguistic intervals, where t_j is the number of linguistic labels in the disjunction j. Some definitions are got for representing formally how the partitions are made.

Definition 1. Lets a set of linguistic labels $SA_j = \{SA_j^1, SA_j^2, ..., SA_j^{i_j}\}$ for a linguistic variable X_j. A *linguistic interval of length c (LI)* is a set of consecutive linguistic labels defined in SA_j, and is represented as: $LI_{j,p}^c = \{SA_j^p, SA_j^{p+1}, ..., SA_j^{p+(c-1)}\}$, where p is the position in SA_j of the first label in the linguistic interval and c is the number of labels in the linguistic interval.

For instance, let us suppose that the set of linguistic labels SA_j is defined over the variable X_j, and SA_j has the linguistic labels: VN, N, NR, P, and VN. The linguistic interval of length 3 $LI_{j,2}^3$ is the set of labels {N, NR, P}.

Definition 2. Let's a real value a_j defined in the domain D_j of the variable X_j, and a linguistic interval of length c $LI_{j,p}^c$, the membership function of $LI_{j,p}^c$ is: $\mu_{LI_{j,p}^c}(a_j) = \sum_{SA_j^z \in LI_{j,p}^c} (\mu_{SA_j^z}(a_j))$, where $z \in [p..c-1]$.

Definition 3. The set of m linguistic intervals defined over m different linguistic variables is known as *Set of m Linguistic Interval (SLI_m)*, and is represented as: $SLI_m = \{LI_{1,p_1}^{c_1}, LI_{2,p_2}^{c_2}, ..., LI_{m,p_m}^{c_m}\}$, where p_j is the position in SA_j of the first linguistic label in the linguistic interval j, and c_j is the number of linguistic labels in the linguistic interval.

So, a antecedent with m disjunctions is represented as a set of m linguistic interval. So, the antecedent of the example anterior is equivalent to: $SLI_3 = \{LI_{1,2}^2, LI_{2,1}^4, LI_{3,4}^2\}$ with $LI_{1,2}^2 = \{SA_1^2, SA_1^3\}$, $LI_{2,1}^4 = \{SA_2^1, SA_2^2, SA_2^3, SA_2^4\}$ and $LI_{3,4}^2 = \{SA_3^4, SA_3^5\}$.

Definition 4. Let's a example $e_i = (x_1^i, ..., x_m^i, y^i, t^i)$ and a set of m linguistic intervals SLI_m, the membership function of SLI_m is:

$\mu_{SLI_m}(e_i) = *(\mu_{LI_{j,p_j}^{c_j}}(x_j))$, where $j \in [1..m]$ and * is a t-norm.

So, the method for obtaining the set SPAR from SLI_m consists in two phases:

1. The $c_j - 1$ possible combinations (Figure 1) are made for each one of the m linguistic intervals $LI_{j,p_j}^{c_j}$ in SLI_m. In general, in each combination p of a linguistic interval j the first linguistic interval is formed for the p first linguistic labels, and the second linguistic interval for the final $c_j - p$ linguistic labels. For each linguistic interval

$LI_{j,p_j}^{c_j}$ the next c_j-1 pairs of linguistic intervals are obtained:

$$(LI_{j,1}^{1} = \{SA_j^1\},\ LI_{j,2}^{c_j-1} = \{SA_j^2, SA_j^3, \ldots, SA_j^{c_j}\})$$

$$(LI_{j,1}^{p} = \{SA_j^1, \ldots, SA_j^p\},\ LI_{j,p+1}^{c_j-p} = \{SA_j^{p+1}, \ldots, SA_j^{p+(c_j-p)}\})$$

$$(LI_{j,1}^{c_j-1} = \{SA_j^1, SA_j^2, \ldots, SA_j^{c_j-1}\},\ LI_{j,c_j}^{1} = \{SA_j^{c_j}\})$$

2. Each pair of linguistic intervals obtained in the anterior phase is converted in a pair of sets of m linguistic intervals. In general, for each combination p of a linguistic interval j in the original SLI_m, the two new sets of m linguistic intervals are equal to SLI_m except in the linguistic interval in the position j that is substituted by the new linguistic intervals $LI_{j,1}^{p}$ and $LI_{j,p+1}^{c_j-p}$ respectively.

2.2 Evaluation Function

For obtaining the evaluation function each SLI_m in SPAR has associated a set of examples E_{SLI_m}, and the a-subsets of examples are defined (Definition 4). The set of examples E_{SLI_m} is: $E_{SLI_m} = \{e_i \in E_R / \mu_{SLI_m}(e_i) > 0\}$, where E_R is the set of examples of the set of linguistic intervals used for obtaining SPAR.

So, the subset E_{SLI_m} is built with the examples of E_R that has a grade of membership to SLI_m over 0.

Definition 5. Let's an output linguistic label SC^k defined in the set of linguistic labels of the consequent SC, and a set of examples E_{SLI_m} of a set of m linguistic intervals SLI_m. $E_{\alpha-SLI_m}^{SC^k}$ is the *a-subset of examples of* E_{SLI_m} *with the output linguistic label* SC_k and is defined as: $E_{\alpha-SLI_m}^{SC^k} = \{e_i \in E_{SLI_m} / \mu_{SC^k}(y^i) > \alpha\}$ with $e_i=(x_1^i, \ldots, x_m^i, y^i, t^i)$.

That is, the set $E_{\alpha-R}^{SC^k}$ is formed by the examples of E_{SLI_m} that have a grade of membership to the output linguistic label SC_k more than a.

For simplify the notation each pair of sets of linguistic intervals (*PSLI*) in SPAR is represented as (SLI_1, SLI_2). Consequently, the subset of examples of SLI_1 and SLI_2 are named E_{SLI_1} and E_{SLI_2} respectively, and the a-subset $E_{\alpha-SLI_i}^{SC^k}$ of examples of E_{SLI_i} with output linguistic label SC_k is denoted as E_i^k where k is the position in SC of the output linguistic label, and i represents SLI_i.

Our proposition for evaluating PSLI is based in the concept of a-subset. Each linguistic interval (SLI_i) in PSLI is evaluated by the function:

$$Eval(SLI_i) = \frac{\sum_{k=1}^{i_y} \left(N(E_i^k) * \left(\sum_{j=1}^{N(E_i^k)} * \left(\mu_{SC^k}(y_k^j), \mu_{SLI_i}(e_{j,k}) \right) \right) \right)}{N(E_{SLI_i})}$$

where i_y is the number of linguistic labels of SC, $N(E_i^k)$ is the number of examples of the a-subset E_i^k, $N(E_{SLI_i})$ is the number of examples of the set of examples E_{SLI_i} associated to the set of linguistic intervals SLI_i, $e_{j,k}$ is the example j in the a-subset of examples of E_i^k with the output linguistic label SC^k, and y_k^j is the output value in the example $e_{j,k}$, and * is a t-norm in the expression $*\left(\mu_{SC^k}(y_k^j), \mu_{SLI_i}(e_{j,k})\right)$.

The next expression is used for evaluating PSLI:

$$\text{Eval(PSLI)} = \sum_{i=1}^{2} Eval(SLI_i)$$

Finally, from the set SPAR is selected the *PSLI* that verifies: PLSI=max(Eval($PSLI_i$)), where i take values since 1 to number of pairs in SPAR.

2.3 Simplifying the linguistic intervals

The two antecedents SLI_1 and SLI_2 of the selected pair of sets of linguistic intervals *PSLI* can be simplified, so the first and last linguistic label for each linguistic interval $LI_{j,p_j}^{c_j}$ in SLI_i is changed; therefore the number of linguistic labels in SLI_i is decremented. The new first linguistic label SA_j^f in each $LI_{j,p_j}^{c_j}$ is the first, which verifies: $\exists e \in E_{SLI_i} / \mu_{SA_j^f}(x_j) > 0$ where x_j is the real value in the position j of the example e, and the last SA_j^l is the one that verifies $\exists e \in E_{SLI_i} / \mu_{SA_j^l}(x_j) > 0$.

Thus, if the first and the last linguistic labels of the simplified linguistic interval are SA_j^f and SA_j^l

respectively, it has l-f+1 linguistic labels, and is represented as $LI_{j,f}^{(l-f+1)}$.

2.4 Create the rules from PSLI

Each SLI_i is equivalent to an antecedent of a rule, so the conversion of this to an antecedent is trivial, an example is shown in the section 2.1. The next equation is used for obtaining the linguistic label of the consequent: $\max_{SC^w} \mu_{SC^w}(v)$, where w=1..$i_y$, and $v = \frac{\sum_{i=1}^{N(E_{SLI_i})} y^i}{N(E_{SLI_i})}$ where E_{SLI_i} is the set of examples of SLI_i, $N(E_{SLI_i})$ is the number of examples in E_{SLI_i}, and y^j is the output real value in the example e_i which belong to E_{SLI_i}.

So, the linguistic label selected for the consequent is the one that has the maximum grade of membership to the medium value of the output values in the examples of E_{SLI_i}.

2.5 End of the algorithm

The algorithm finishes when for each rule R of the model verify: $\forall e_i \in E_{SLI}, *(\mu_{SLI}(e_i), \mu_{SC^w}(y^i)) >= 0$, with $e_i=(x_1^i, ..., x_m^i, y^i, t^i)$, SLI is the set of linguistic intervals equivalent to the antecedent of R, and SC^w is the label of the consequent of R.

3 Proof of the algorithm

The algorithm has been applied to obtain a linguistic model of the human's walk. Human's walk has been studied using biomechanical models [3,5,7]. There has been some debate as to the most appropriate method of defining joint angles [5], which are a rotation of the distal segment relative to the proximal segment. One of these rotations is defined as flexion and extension (FE). The FE angles of the hip (RHIP), knee (RKNEE) and foot (RDORSI) of the right leg are used as input variables, as output variable is used the FE angle of the knee of the left leg. A set of seven linguistic labels is defined over input and output variables. The linguistic labels are: Very Negative(VN), Negative(N), Few Negative(FN), Norm(NR), Few Positive(FP), Positive(P) and Very Positive(VP). The set of examples is property of Dr. R. Baker and are obtained from the web "guardian.curtin.edu.au/cga/data/index.html". Table 1 shows the model, DHIP, DKNEE, DDORSI and IKNEE are the linguistic intervals obtained.

Table 1. Obtained model

Rule	DHIP	DKNEE	DDORSI	IKNEE	Rule	DHIP	DKNEE	DDORSI	IKNEE
1	MN	MN to N	P to MP	MN	11	PN	MN	MP	PP
2	MN	N to PN	NR to PP	N	12	NR	MN	MP	MP
3	P to MP	NR to MP	NR to PP	MN	13	MP	N	NR	PN
4	MN to PP	PN to MP	N to PN	N	14	MP	MN	NR	MN
5	MP	MN	PP	MN	15	MP	N	P	MN
6	NR to P	N	P to MP	MP	16	MP	N	P	PN
7	P	PN	P	MP	17	P	PN	PP	P
8	MN to PN	NR to MP	MN	PN	18	MP	PN	PP	NR
9	NR	MP	MN	N	19	MP	PN	P	PN
10	N	MN	MP	PN	20	MP	NR	P	MN

4 Conclusions and futures works

A new algorithm for inducing a system of linguistic rules from a group of data is defined. The obtained results allow to confirm that the error made by the used methodology is similar to classic algorithms as ID3 and CART, and new algorithms like ADRI [2]. It is of highlighting that in front of these algorithms the use of linguistic variables from the origin of the same one, reinforces the descriptive and qualitative character of the obtained pattern; in front of other quantitative methods. The work of immediate future that we are making is to incorporate a refinement phase by means of the use of technical as ANFIS, or genetic algorithms that allow to diminish the errors of the pattern.

Acknowledgements
This work has been financed by the project TIC2000-1362-C02-02 of the Ministry of Science and Technology of Spanish state.

References

[1] Breiman L., Friedman J., Olshen R., and Stone C. "Classification and regression tree". *Monterey, Ca.:Wadsworth*, 1984.

[2] Jiménez L. "Modelización Difusa de Sistemas mediante Técnicas Inductivas", *Tesis Doctoral. Universidad de Granada*.1997.

[3] Luttgens K., Wells K. F.. "Kinesiología: Bases científicas del movimiento humano". *CBS College Publishing*. 1985.

[4] Quinlan J. R. *"Induction of decision tree"*, Machine Learning, 1, 81-106.

[5] Vaughan C.L. "Dynamics of Human Gait". *Human Kinetics Publishers*. 1992.

[6] Zadeh L.A. "The concept of a linguistic variable and its applications to Approximate reasoning (Part I, II, III) ". *Information Sciences*, 1975.

[7] Zatsiorsky V. M. "Kinematics of Human Motion". *Human Kinetics*. 1998.